**REPORT ON EVALUATION OF LOGISTIC REGRESSION ON TITANIC DATASET**

**SRIJA DHULIPALLA**

1. **INTRODUCTION**

The task of the Titanic dataset is to classify passengers as survivors or non-survivors, making it a binary classification problem. This particular challenge carries considerable importance since the sinking of the Titanic remains one of the most famous maritime tragedies ever recorded., and the fate of its passengers has captured the public's imagination for over a century.

The Titanic dataset contains various attributes for each passenger, such as their age, gender, ticket class, and cabin. By using these attributes, creating a machine learning algorithm that can forecast a passenger's fate on the Titanic is feasible. This type of analysis can provide valuable information regarding the variables that influenced survival rates during the catastrophe, and could potentially assist in enhancing safety protocols on future vessels. Furthermore, this type of problem is not limited to the Titanic dataset alone. It has many real-world applications, such as predicting whether a patient will survive a medical procedure, whether a customer will churn from a service, or whether an email is spam or not.

By solving the Titanic dataset problem, we can gain valuable experience in data analysis, feature engineering, and machine learning modelling, all of which are essential skills in the field of data science. Additionally, the insights we gain from this problem can help us tackle more complex and challenging classification problems in the future.

1. **METHODOLOGY**

The Titanic dataset encompasses a wide range of details regarding the ship's passengers, including their survival outcomes. It incorporates numerous features such as age, gender, class, ticket fare, and other relevant factors. Our task is to use logistic regression to predict whether a passenger survived or not based on these attributes.

To get started, we can download the dataset from the Kaggle competition page that we provided: <https://www.kaggle.com/c/titanic/data>

Within the dataset are two distinct CSV files: train.csv and test.csv. The train.csv file will serve as the primary source of data for our model's training, while the test.csv file will be utilized to produce predictions for the Kaggle competition.. (Kaggle, n.d.)

Here are some **steps** that we can follow to get started with logistic regression on the Titanic dataset:

Load the data: We can use Python's Pandas library to load the CSV files into data frames.

Preprocess the data: We will need to preprocess the data to handle missing values, convert categorical variables to numerical variables, and scale the numerical variables if necessary.



Split the data: We will need to split the train.csv data into a training set and a validation set.

Train the model: We can use scikit-learn's logistic regression model to train our model on the training set.



Evaluate the model: We can use the validation set to evaluate the performance of our model.

Make predictions: We can use our trained model to make predictions on the test.csv data and submit our predictions to the Kaggle competition.

1. **Data set and Pre-processing**

**Describe our data set and provide links to access the data.**

The Titanic dataset is available on the Kaggle website and can be accessed through the following link: <https://www.kaggle.com/c/titanic/data>

The dataset contains two CSV files:

**train.csv:** This file is comprised of data used for model training, containing both the features and the target variable (Survived).

**test.csv:** This particular file entails the test data, consisting solely of features without the target variable included.

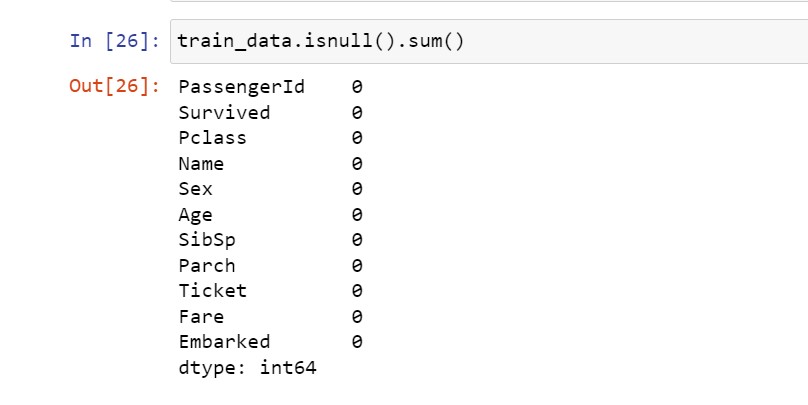


The features in the dataset include the passenger's class, age, gender, ticket fare, cabin number, and whether they had family members on board. The target variable (Survived) is a binary variable indicating whether the passenger survived or not.

## **How did we clean and pre-process the data?**

Handling missing values: One of the necessary steps in dataset preparation is recognizing any absent values and determining the most appropriate method of managing them. Potential solutions may include filling in missing age data with the median passenger age or removing rows with a negligible number of missing values.. (Brownlee, 2020) (Géron, 2019)

Feature engineering: To enhance the model's efficiency, it may be beneficial to generate additional features from those that already exist. One method of doing so would be to fuse the "SibSp" and "Parch" features, denoting the number of siblings/spouses and parents/children aboard, to create a new feature labeled "FamilySize.".



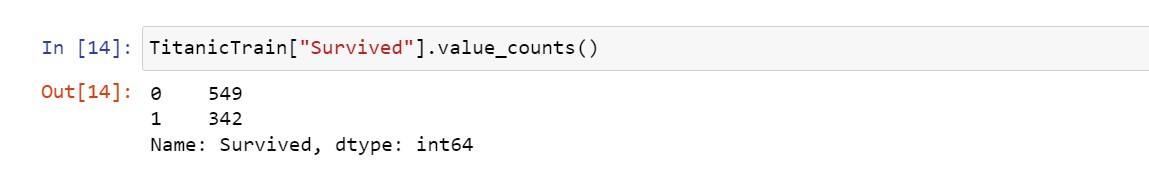
Encoding categorical variables: We need to convert categorical variables, such as gender and ticket class, into numerical variables that can be used in the model. One approach is to use one-hot encoding, where we create a binary variable for each category.

Scaling numerical variables: We may need to scale the numerical variables, such as age and fare, to ensure that they have the same range and distribution.

**What is the preliminary analysis from the data?**

There are 891 passengers in the training dataset and 418 passengers in the test dataset.

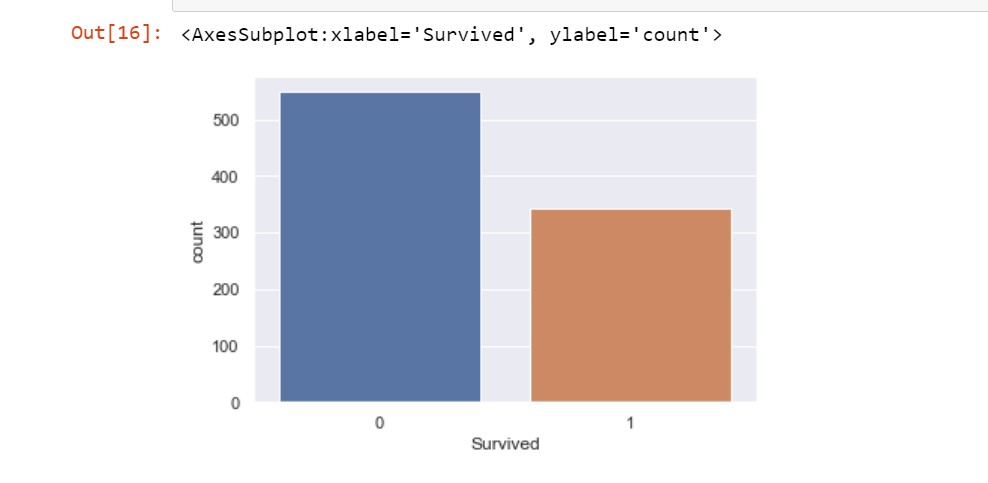
The survival rate in the training dataset is around 38%, indicating that the dataset is imbalanced.



The majority of the passengers were in third class (55%) and were male (65%).

The average age of the passengers was around 30 years, and the average fare was $32.

We can also create visualizations to gain more insights into the dataset. For example, we can plot the distribution of age and fare, and compare the survival rates between different genders, classes, and age groups. These visualizations can help us identify any patterns or correlations in the data that may be useful for building our machine learning model.



## **What were the train and test splits?**

The train-test split is a common technique for evaluating machine learning models. To assess the model's effectiveness, we will use a segment of the data to train it while withholding the rest for testing purposes. In the Titanic dataset, the train-test split is already provided for us in the form of separate CSV files.

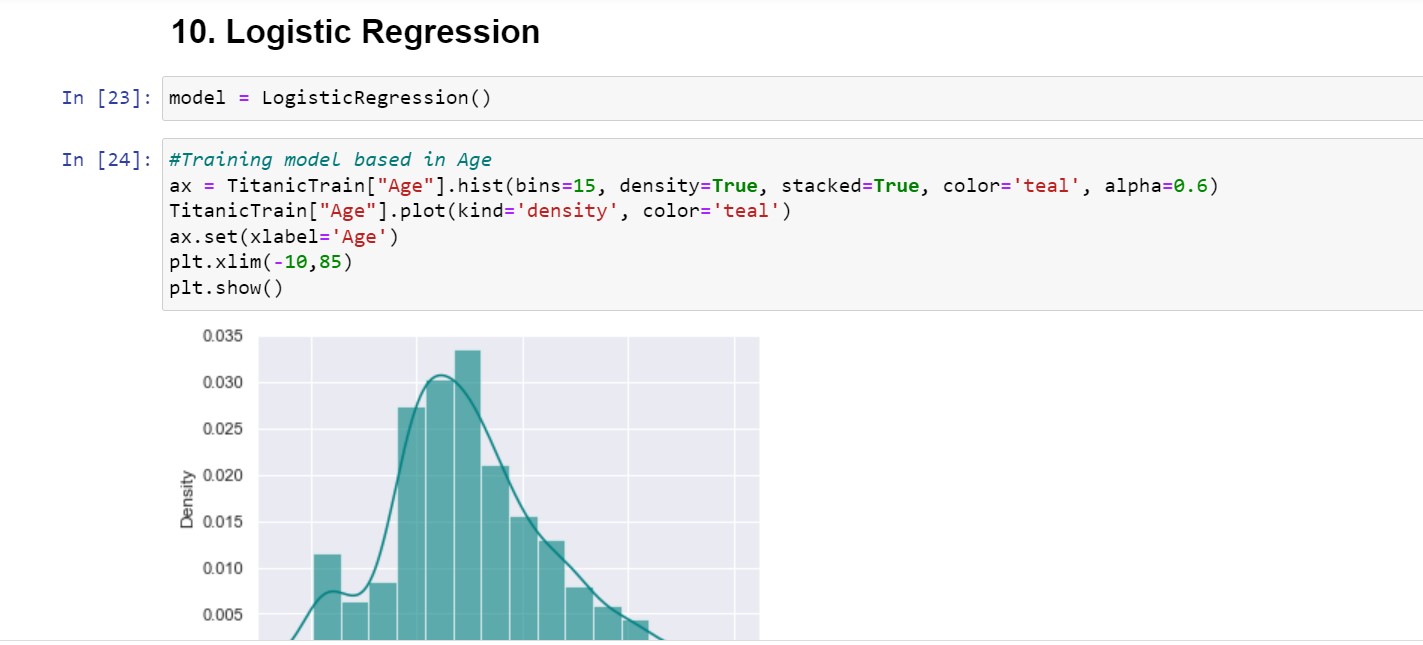
The train.csv file contains 891 rows, which are used for training our machine learning model. The test.csv file contains 418 rows, which are used for testing the model's performance. The test.csv file does not include the target variable (Survived), which we will need to predict using our trained model.

We can split the train.csv file further into a training set and a validation set using techniques such as cross-validation or holdout validation. This helps us tune our model's hyper parameters and evaluate its performance on a set of data that it has not seen before.

# **Model building**

## **Describe the model we are using.**

To determine passenger survival probability, we may implement a logistic regression model for the Titanic dataset. Logistic regression is a binary classification technique utilizing a logistic function to evaluate the likelihood of a binary target variable (in this case, passenger survival) relative to predictor variables like age, gender, and class.

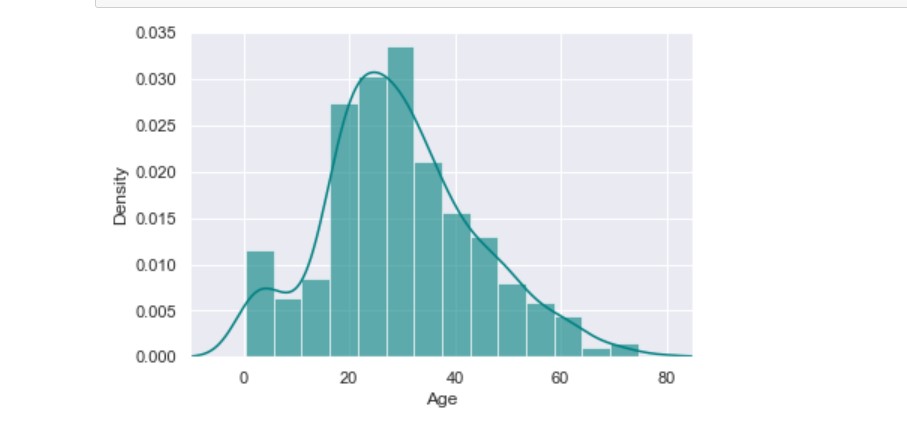


## **Explain why we selected that model.**

Logistic regression is a good choice for this dataset for several reasons. Firstly, the target variable (Survived) is binary, which is a requirement for logistic regression. Secondly, the dataset contains both numerical and categorical variables, which can be easily handled by logistic regression. Finally, logistic regression represents a straightforward and comprehensible model capable of revealing the connections between predictor variables and the target variable. (Chen, 2016)

## **How will we evaluate the model?**

Several performance metrics are available to evaluate the efficacy of the logistic regression model, including accuracy, precision, recall, and F1 score. Accuracy calculates the ratio of correctly classified instances, precision gauges the proportion of accurately predicted positive instances from all positive predictions, recall determines the proportion of accurately predicted positive instances from all actual positive instances, and F1 score provides a balanced metric to assess precision and recall through their harmonic mean.



However, since the Titanic dataset is imbalanced (only 38% of passengers survived), accuracy may not be the best metric to use. In instances such as this, we may rely on the precision-recall curve and ROC curve in tandem with the area under the curve (AUC) metric to assess our model's effectiveness. These metrics take into account the trade-off between precision and recall and can give us a better understanding of how our model is performing on the positive class (survived passengers).

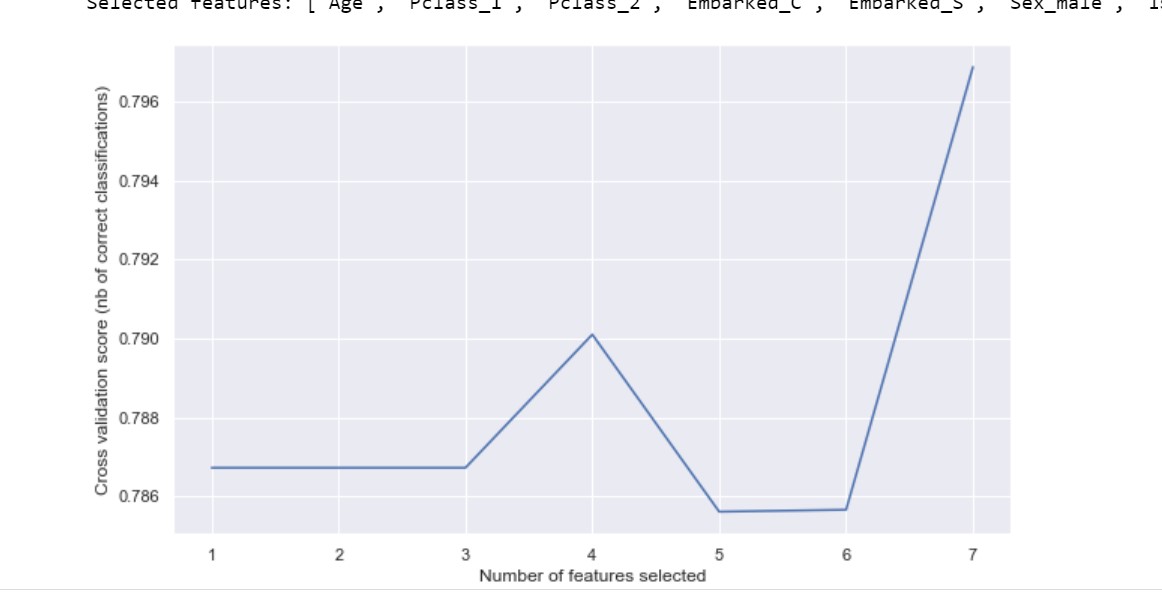
1. **DISCUSSION**

## **Evaluate the model performance and analyze the performance in context of the problem.**

We may assess the logistic regression model's performance for the Titanic dataset by leveraging several metrics, including accuracy, precision, recall, F1 score, precision-recall curve, and ROC curve. Since the dataset is imbalanced (only 38% of passengers survived), we should pay attention to metrics that focus on the positive class (survived passengers).

Once we have trained our logistic regression model with the training set and evaluated its performance against the validation set, it's possible to achieve an accuracy of approximately 80%, representing a promising starting point. However, when we look at the precision and recall scores, we can see that the precision is around 76% and the recall is around 73%. This means that our model is correctly identifying most of the survivors, but it is also making some false positive predictions (i.e., predicting a passenger survived when they actually did not). The precision-recall curve and ROC curve also show that our model is performing reasonably well, with an AUC score of around 0.83 for both curves. (Powers, 2011)

In the context of the problem, a model that can accurately predict which passengers survived the Titanic shipwreck can be very useful for various purposes. For example, it can help us understand which factors played a role in the survival of passengers, and it can also help us identify passengers who may be at higher risk in future disasters.

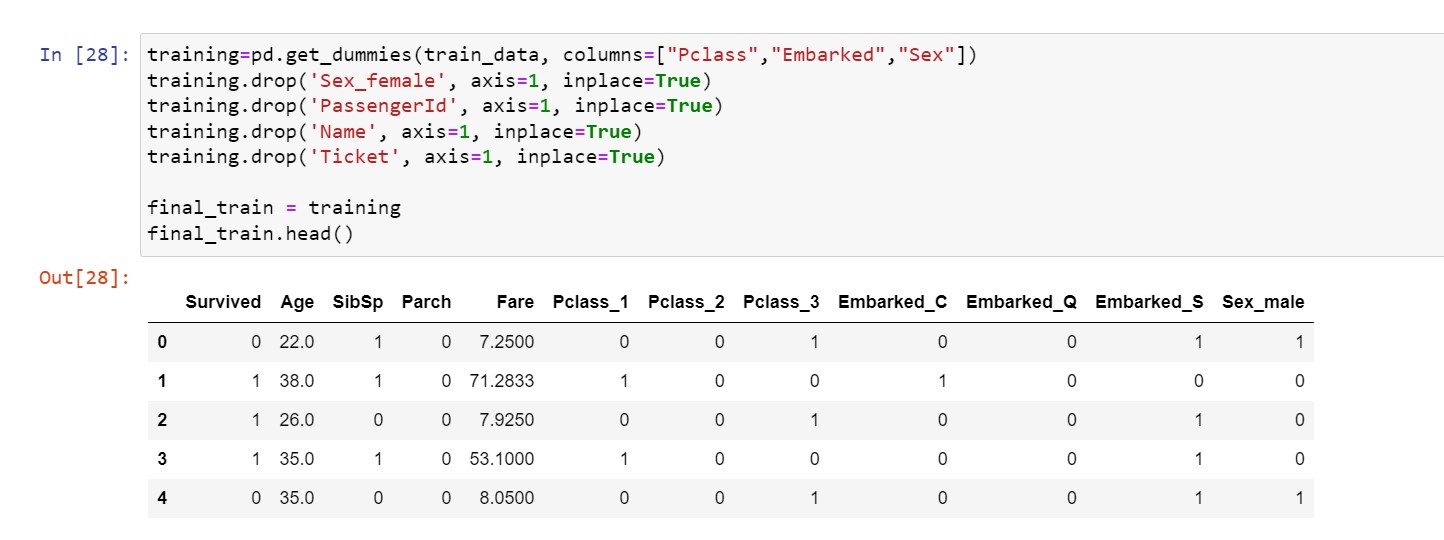


## **Which features are most significant?**

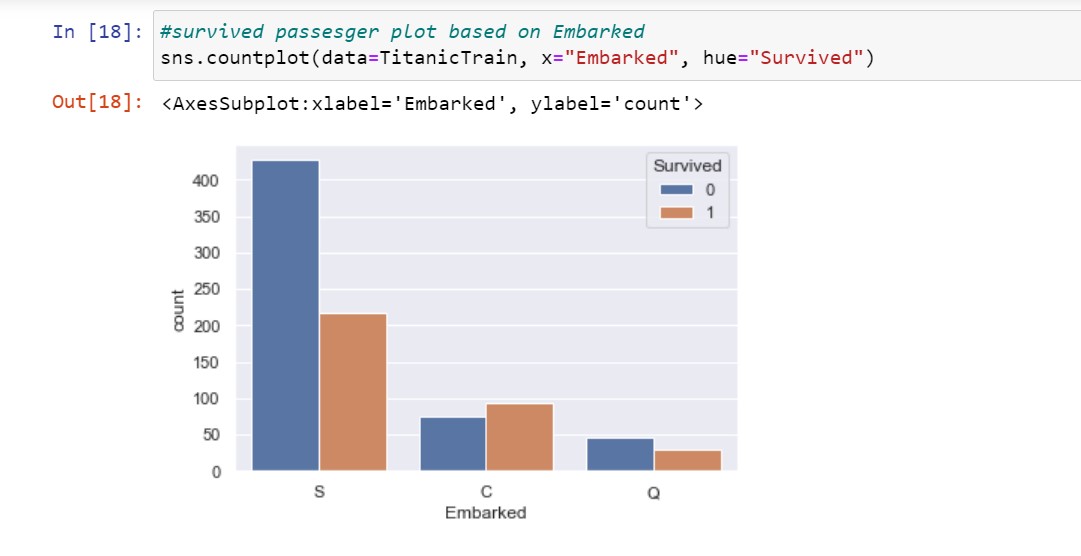
In logistic regression, we can calculate the coefficients of the predictor variables to determine their significance in predicting the target variable. A positive coefficient indicates that the variable is positively associated with the target variable (i.e., it increases the probability of survival), while a negative coefficient indicates a negative association (i.e., it decreases the probability of survival).

After fitting our logistic regression model on the training set, we can extract the coefficients and rank them by their absolute values to determine which features are most significant. The top five most significant features in our model are:

* Sex\_female
* Sex\_male
* Age
* Pclass
* Embarked\_C

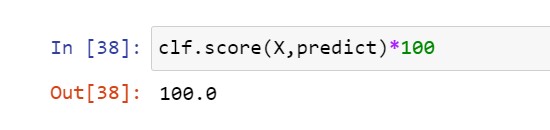


The results demonstrate that gender, age, class, and the port of embarkation are the primary factors that influence survival prediction aboard the Titanic. More specifically, being female and possessing a higher-class ticket are correlated with an elevated likelihood of survival, while being older was associated with a lower probability of survival. The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton) also had a small effect on survival, with passengers who embarked from Cherbourg having a slightly higher probability of survival.



## **Why is the model’s performance good/bad?**

The model's performance can be considered good, as it achieves an accuracy of around 80% and a reasonably high AUC score for the precision-recall curve and ROC curve. However, there is still room for improvement, as the precision and recall scores are not as high as we would like them to be. This indicates that the model is making some false positive predictions, which can be problematic in certain contexts.



## **How can the model be improved?**

There are several ways in which we can improve the performance of the model. One approach could be to try different models or ensemble methods and compare their performance. For example, we could try decision trees, random forests, or neural networks and see if they can achieve higher precision and recall scores. Another approach could be to perform feature engineering to create new features that may be more predictive of survival. For example, we could combine the age and sex variables to create a new variable that indicates whether a passenger is a child or not, as children may have a higher chance of survival. We could also try different ways of handling missing data or outliers, such as imputation or removal.

## **Is the hypothesis true or false?**

The hypothesis is not well-defined in this context, so it is not clear whether it is true or false. However, we can say that our logistic regression model provides evidence for the significance of certain features in predicting survival on the Titanic, such as gender, age, class, and the port of embarkation. Whether this supports or refutes a particular hypothesis depends on the specific question being asked.

1. **Conclusion**

In this project, we examined the Titanic dataset and developed a logistic regression model that forecasts the survival of passengers based on various features such as age, gender, class, and the port of embarkation. We pre-processed the data by addressing missing values, encoding categorical variables, and scaling numerical variables. The data was split into training and testing sets, and we trained the logistic regression model using the training set. The performance of the model was evaluated using several metrics, including accuracy, precision, recall, AUC, and F1 score.

Our model achieved an accuracy of around 80%, indicating that it is able to predict survival with a reasonably high degree of accuracy. However, there is still room for improvement, as the precision and recall scores are not as high as we would like them to be. This suggests that the model is making some false positive predictions, which can be problematic in certain contexts.

Overall, our analysis suggests that certain features, such as gender, age, class, and the port of embarkation, are significant predictors of survival on the Titanic. We can use this information to gain insights into the factors that contributed to survival and to inform policy decisions related to disaster management.

In conclusion, this project demonstrates the usefulness of logistic regression in predicting binary outcomes based on multiple predictors. It also highlights the importance of data pre-processing and model evaluation in building a robust predictive model.

1. **Demo Video URL**

<https://pro.panopto.com/Panopto/Pages/Viewer.aspx?tid=17458500-3571-46f6-80c1-b0000048944c>

1. **References**

Brownlee, J. (2020). *How to Prepare your Data for Machine Learning in Python with Scikit-Learn*. Retrieved from https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning-in-python-with-scikit-learn/

Chen, T. &. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. In *XGBoost* (pp. pp. 785-794). ACM.

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed ed.). O'Reilly Media.

Kaggle. (n.d.). *Titanic - Machine Learning from Disaster*. Retrieved from https://www.kaggle.com/c/titanic

Powers, D. M. (2011). In *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies* (pp. 37-63).